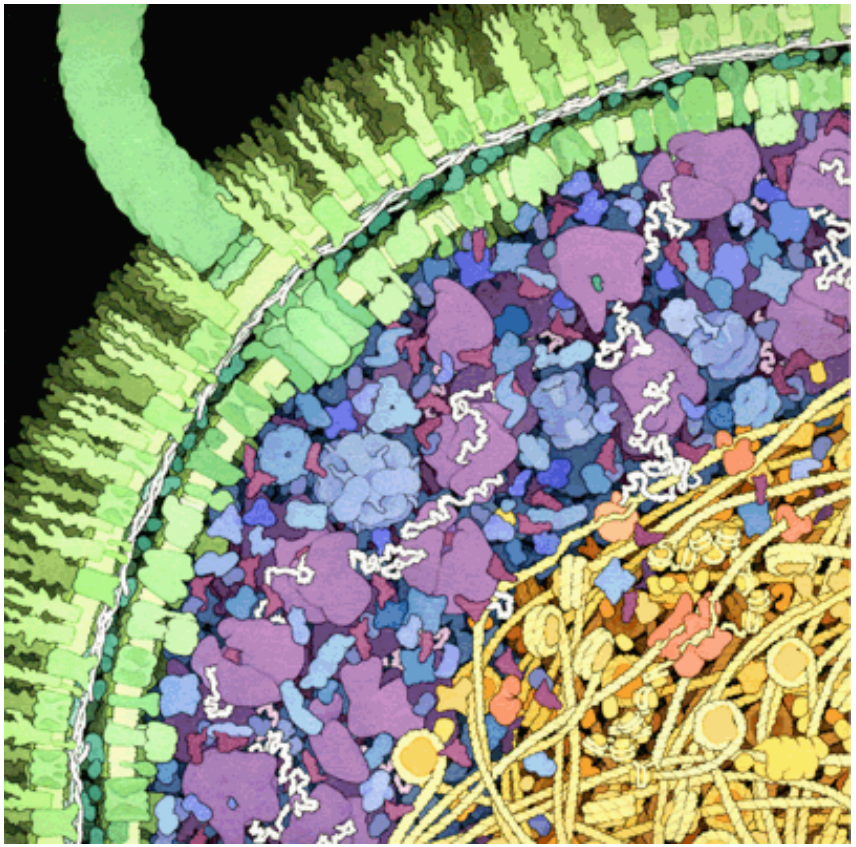
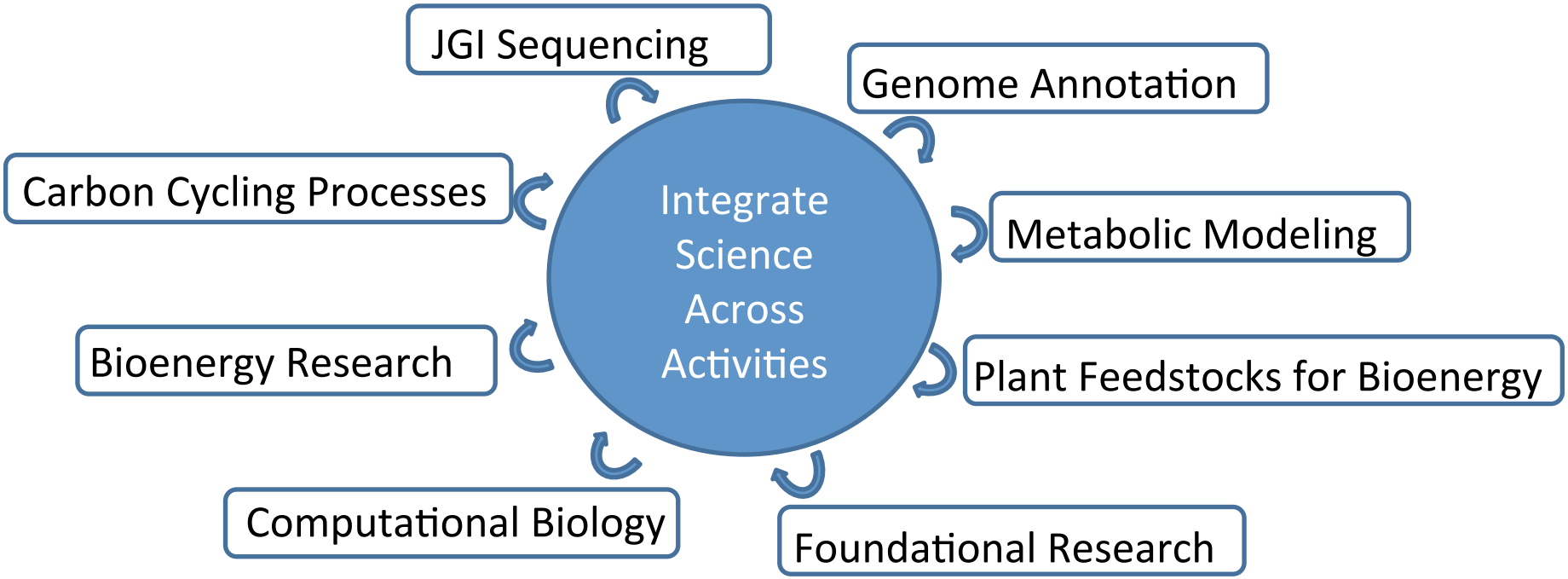
KBASE
predictive biology
DOE Systems Biology Knowledgebase

Building the Systems Biology Knowledgebase

Tom Brettin
Oak Ridge National Laboratory
brettints@ornl.gov
outreach@kbase.us
kbase-users@lists.kbase.us
kbase-devel@lists.kbase.us



Integrate science and the science community

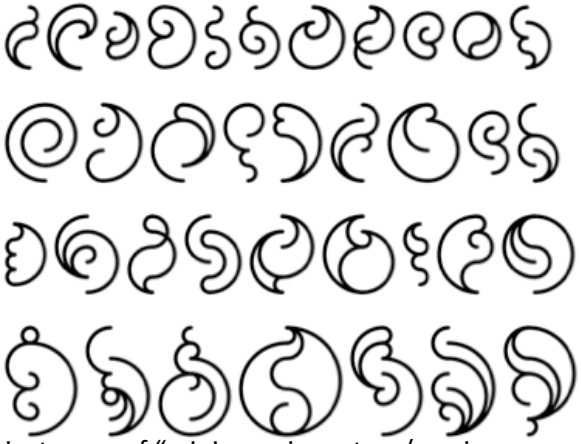


There is a tremendous wealth of data and information in the Genomic Sciences program. The Knowledgebase (Kbase) is an opportunity to integrate this data and information both within individual activities as well as to integrate together different activities.

Everyone should be a contributor!

KBASE:

A. Professional Computational Biologists
B. Data generators and basic analysts
C. Knowledge Seekers
D. Knowledge Generators

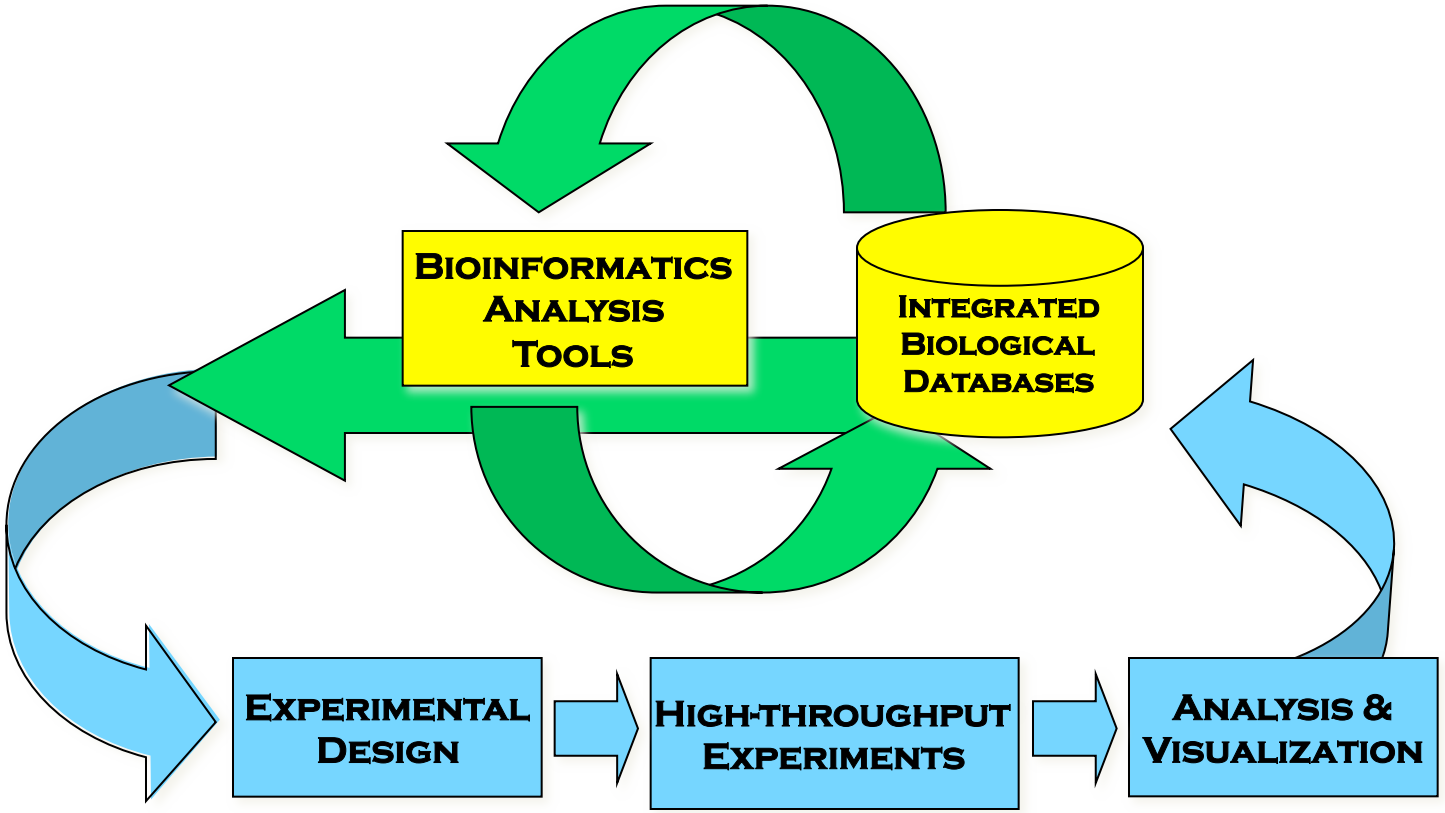


instances of "minimum inventory/maximum diversity" systems, a term coined by Peter Pearce in his book, Structure in Nature Is a Strategy for Design (MIT Press, 1978).

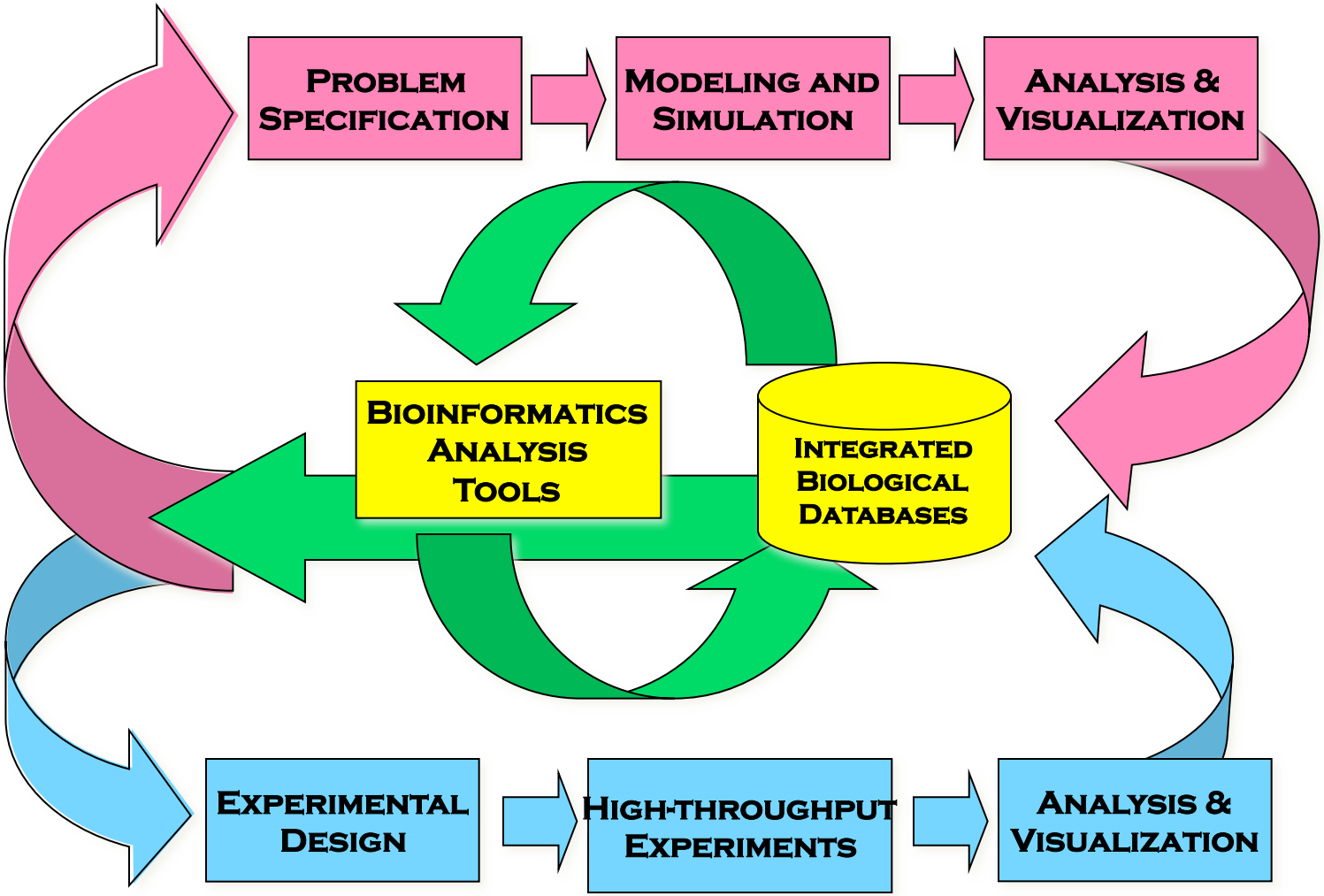
Therefore we aim to:

- Create a powerful framework for programmatic access to data and functions of Kbase. (Users A,B)
  - Ultimately provide stubs for use in PERL, PYTHON, R, MATLAB, Galaxy, etc.
- Create a set of packaged "Widgets" that make placement and recognizable display of Kbase "functions" on web pages (or within perhaps other apps), easy and identifiable. (Users B)
- Create a "simplified" portal for search and aggregation of data for data consumers and Knowledge Seekers. (Users C,D)
- *Create a innovative platform for knowledge creation, evolution and sharing.*

An Integrated View of Modeling, Simulation, Experiment, and Bioinformatics



An Integrated View of Modeling, Simulation, Experiment, and Bioinformatics

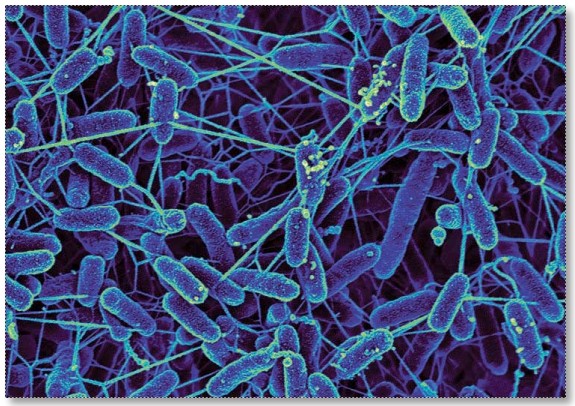


KBASE
predictive biology
DOE Systems Biology Knowledgebase

Systems Biology Knowledge

Knowledgebase enabling ***predictive*** systems biology.

- Powerful modeling framework.
- **Community-driven**, extensible and scalable **open-source** software and application system.
- Infrastructure for integration and reconciliation of algorithms and data sources.
- Framework for standardization, search, and association of data.
- Enable model based **experimental design** and **interpretation** of results.



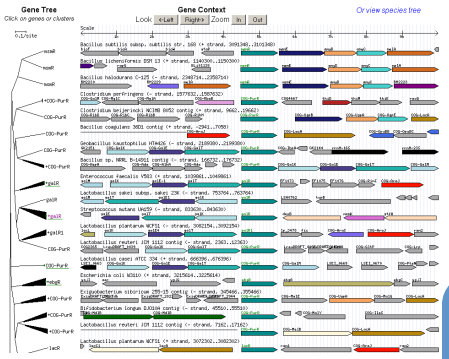
Microbes

Communities

Plants

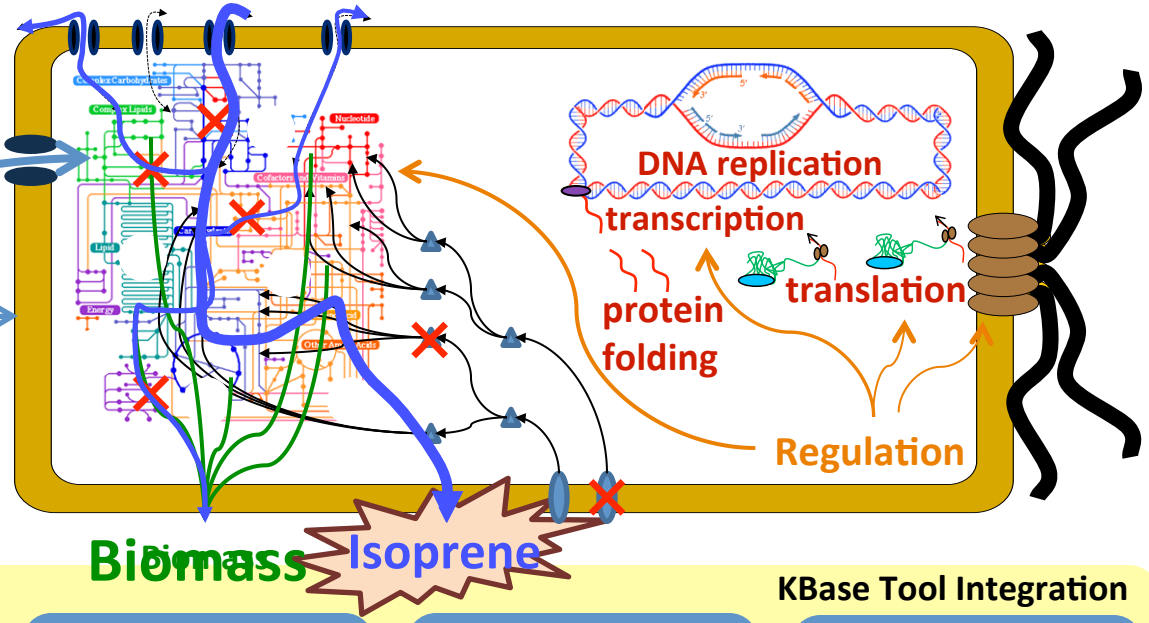
Engineering a Microbe for Biofuel Production

Annotated Genome



Feed Stock

Stresses
Hydrolysate,
pH, Salt,
End product,
intermediates



KBase Tool Integration

Annotation algorithms

Metabolic model generation

Regulatory network inference

Fitting kinetic model parameters

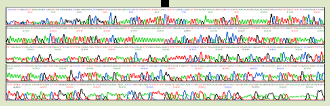
Proposing strain optimizations

Metabolic reconstruction

Model optimization algorithms

Other functional modeling

Predicting pathway fluxes



Genome Sequence

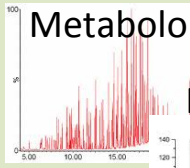
Comparative Genomics

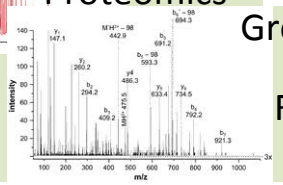
KEGG
Brenda
BioCyc
Published models

BIOLOG

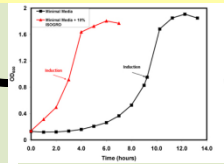
Gene KO Phenotypes

Transcriptomics

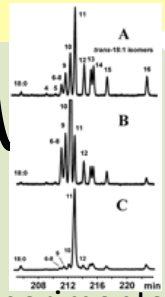




Proteomics



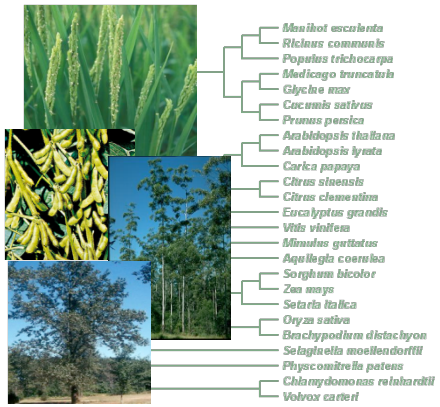
Growth curves

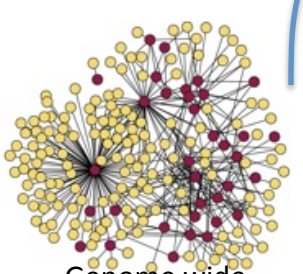


Flux tracing experiments

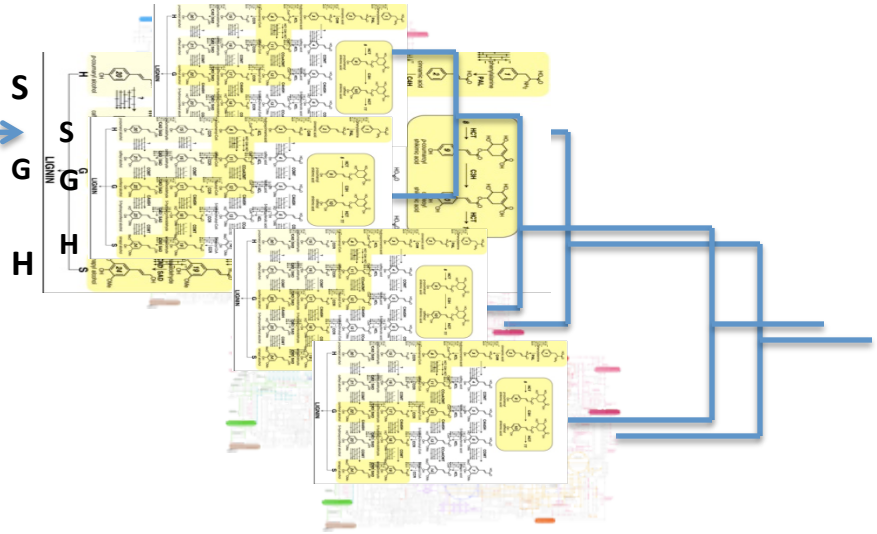
KBase Data Integration

Modifying Lignin Biosynthesis





Genome wide
Correlative analysis



SNPs3D

PolyPhen-2

SNP influenced changes in protein structure and function

Pathway predictions

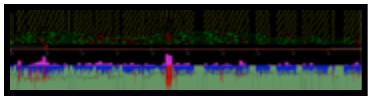
- Model optimization
- validation

Plant systems modification

- Genome annotation algorithms
- Comparative genomics

- Network inference
- Pathway reconstruction
- Omics & SNP overlay

Phylogenomics
Modeling phase I



phytozome

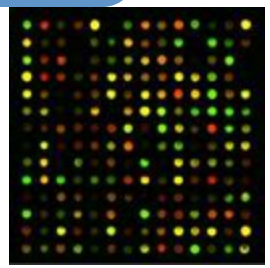
NCBI

Phenotype
Mutant population
Resequencing data

iPlant Collaborative

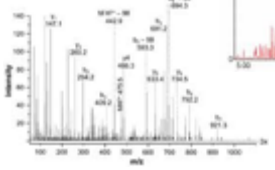
BESC BioEnergy Science Center

jbei Joint BioEnergy Institute

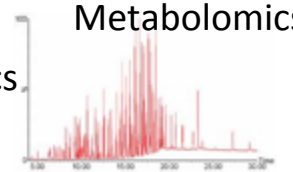


Transcriptomics

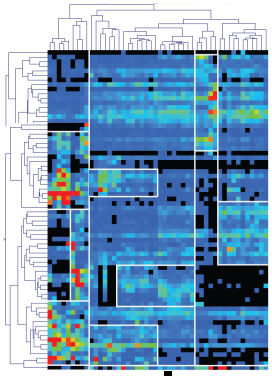
Proteomics



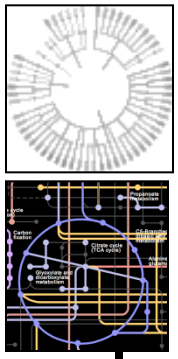
Metabolomics



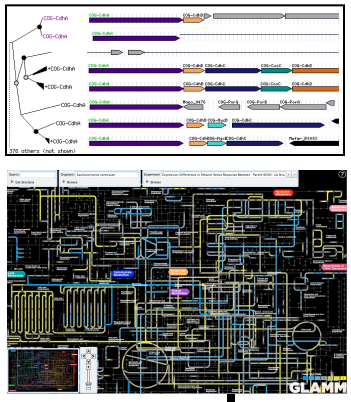
Culturing Recalcitrant Microbes from Communities



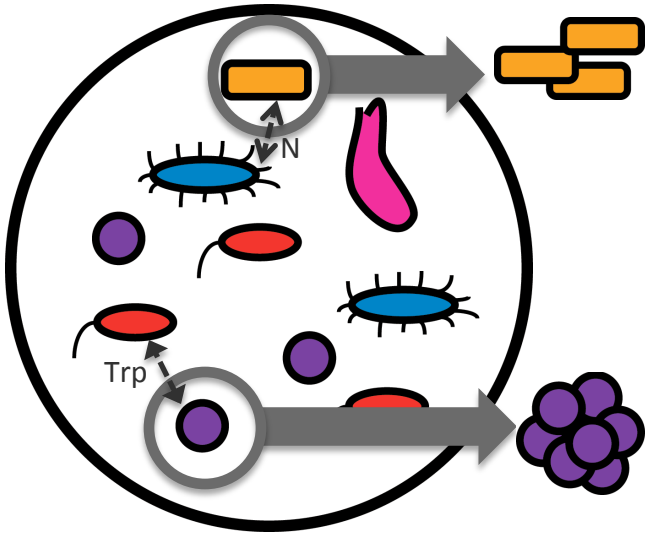
Covariation Analysis, Phylogenetically and Functionally Interesting Keystone Species



Phylogenetic Inference



Gene Functional Annotation



Differential Gene Expression

Population Statistics

Isolate Genomes and Models

Annotation and Metabolic Reconstruction

Predict Syntrophic Interactions

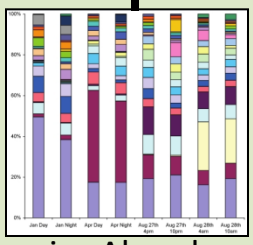
Isolate vs. Community Phenotype

Comparative Metagenomics

Genome Assembly from Metagenomics

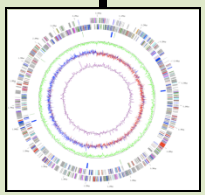
Regulation and Functional Modeling

Predict Culturing Conditions



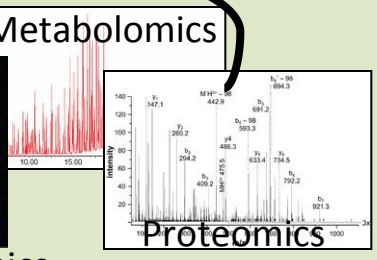
Species Abundance

Functional Gene Abundance



Phylo binning and scaffolding

Transcriptomics



Temp
pH
Salinity
Amino Acids
Cofactors
Syntrophies

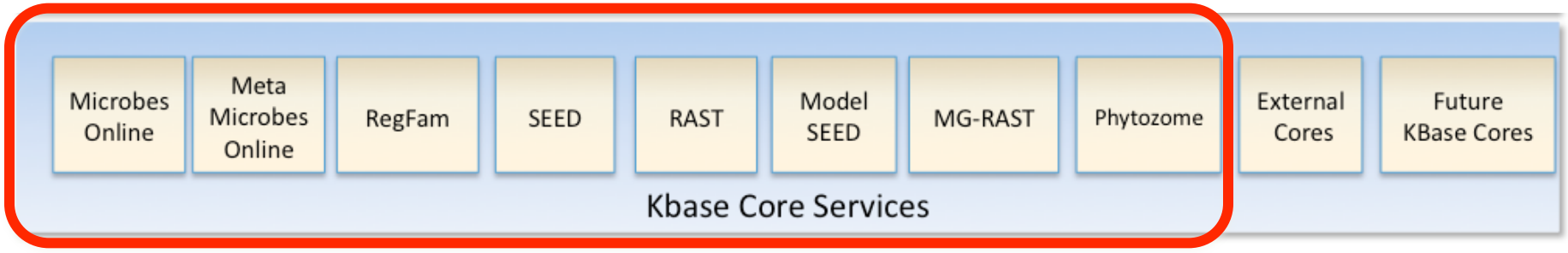
What the KBase Needs To Provide?

- Scalable compute and data capabilities beyond that available locally
- Distributed infrastructure available 24x7 worldwide
- Integration with local bioinfo systems for seamless computing and data management
- Enables leverage of remote systems administration and support via service providers
- Enables access to state of the art facilities at fraction of the cost (SPs just add more servers)
- Centralized support of tools and data
- Bottom line ⇒ <u>enable biologists to focus on biology</u>

Leverage Existing Investments

- We leverage the considerable investments in existing integrated databases and analysis environments

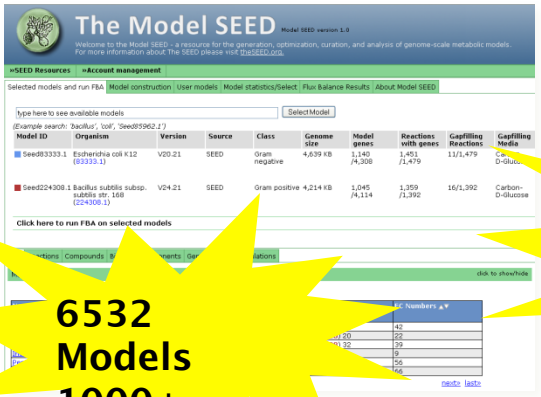
- Key challenge: How we build on these systems yet provide to the community an integrated view for future development



Microbes Online

1000s Data Sets
300+ Daily Users

Model SEED



6532 Models
1000+ Users

MG-RAST

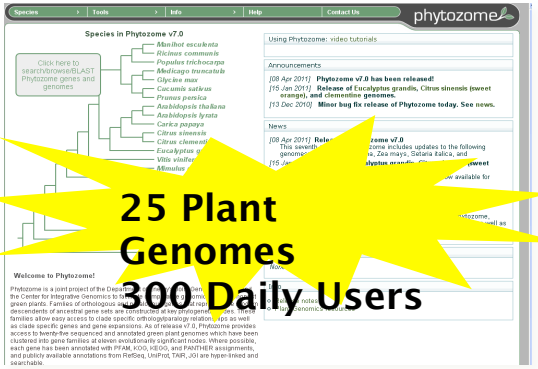
41,000 Metagenomes
500+ Daily Users

Meta Microbes Online

153 Metagenomes
100+ Daily Users

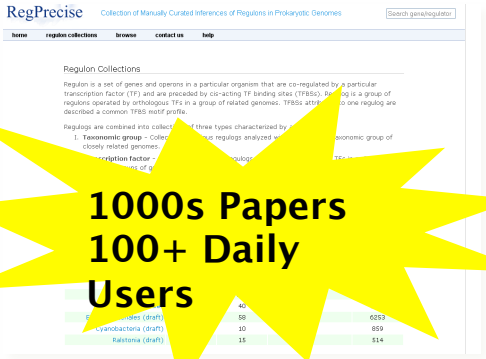
20,000+ users

Phyotozome



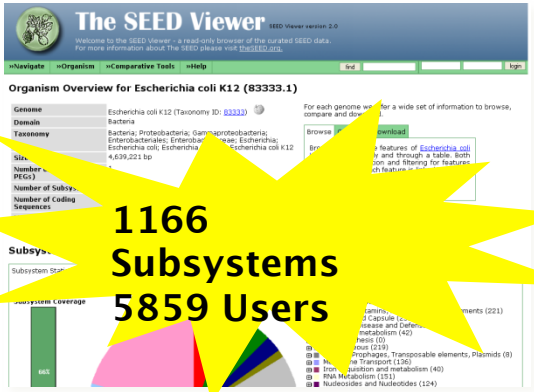
25 Plant Genomes
300 Daily Users

RegFam



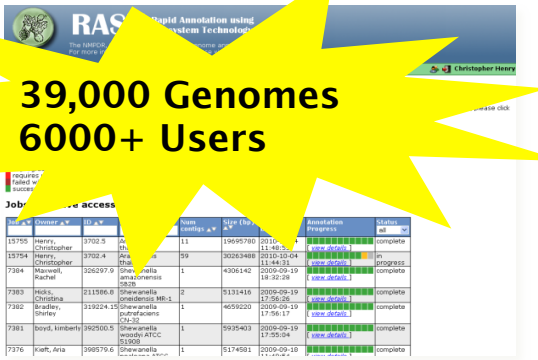
1000s Papers
100+ Daily Users

The SEED

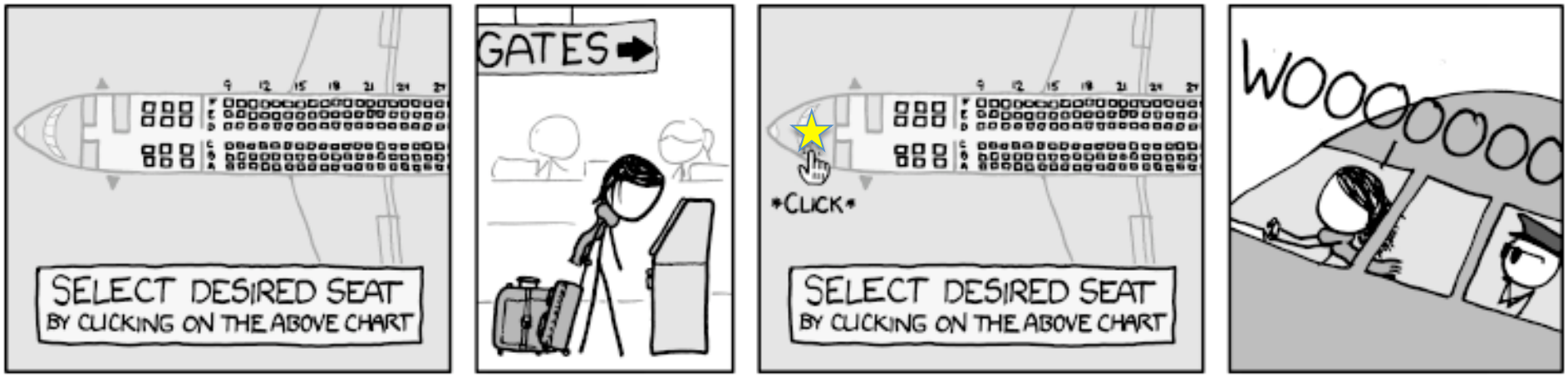


1166 Subsystems
5859 Users

RAST



39,000 Genomes
6000+ Users



Our vision is to put users in the drivers seat.

DOE Systems Biology Knowledgebase

KBASE

Data and modeling for predictive biology

Overview of Infrastructure

Tom Brettin and Rick Stevens
Oak Ridge and Argonne
National Laboratories

U.S. DEPARTMENT OF ENERGY | Office of Science

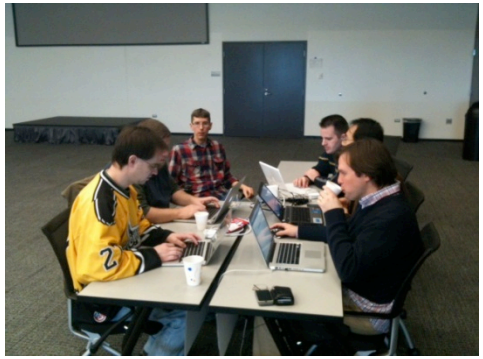
KBASE
predictive biology
DOE Systems Biology Knowledgebase

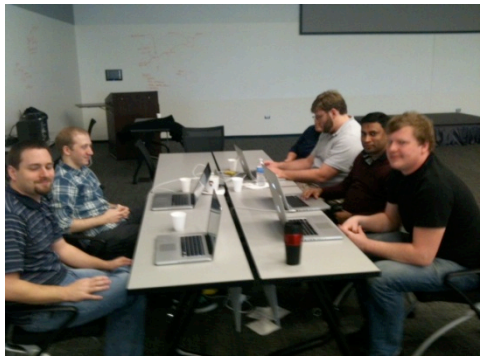
Working As One Team

Communities Hackathon
Jan 2012, LBL

Plant CDM Design
and Build
Jan 2012, ORNL

First Internal Kbase Build – Feb 2012, ANL

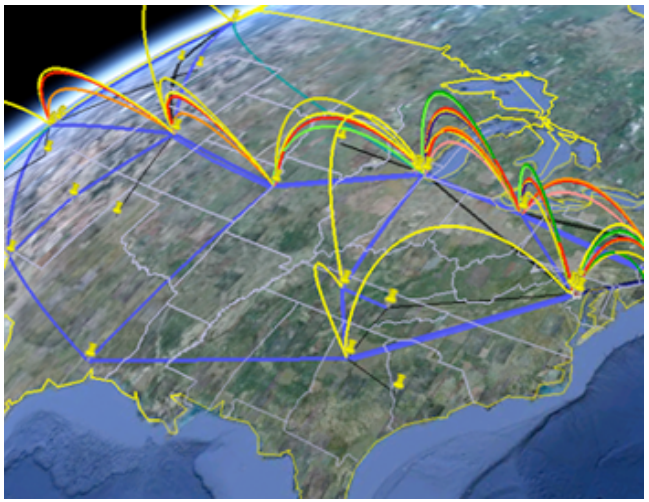




Scientific Software Technical Reviews (May 2-3, 2012)

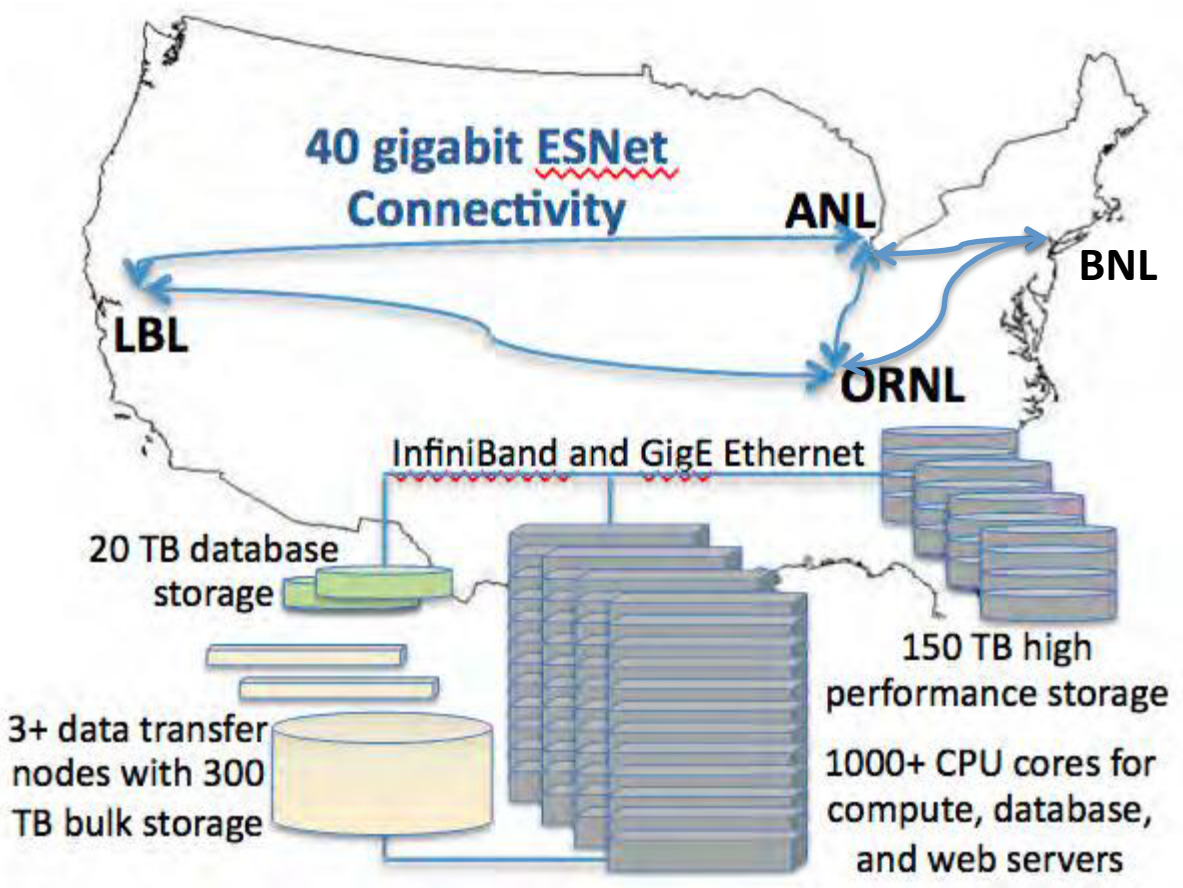
KBASE predictive biology
DOE Systems Biology Knowledgebase

Energy Sciences Network (ESnet)



- **ESnet backbone (ESnet4) is a national 10 Gbps optical circuit infrastructure**
- **ESnet shares its optical network with Internet2**
- **ESnet's IP network functions as a Tier 1 internet service provider**

KBase leverages ESNet for 10+ Gb/s data transfer between all nodes



KBASE
predictive biology
DOE Systems Biology Knowledgebase

The DOE KBase Cloud

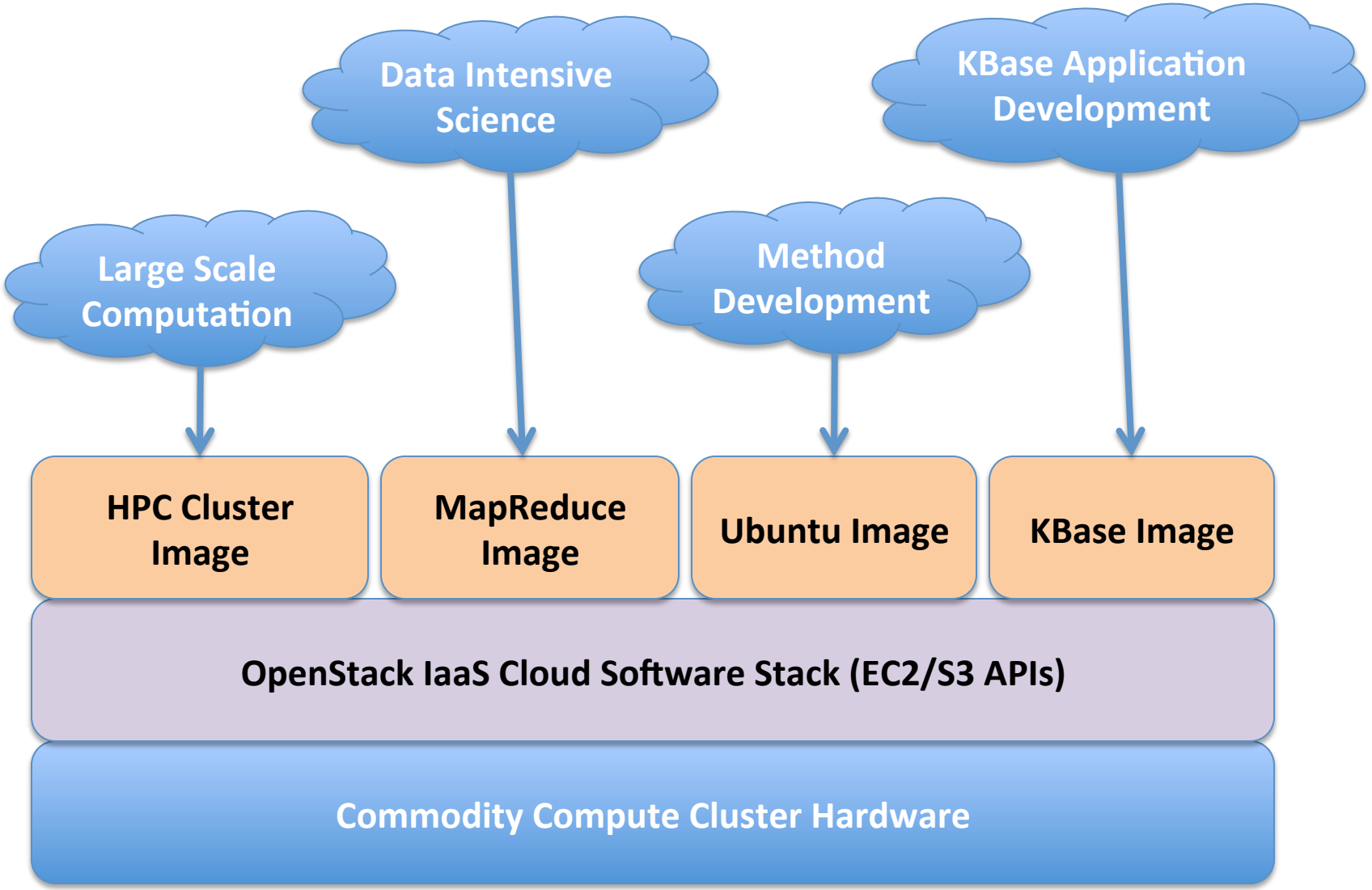
Built on the DOE ASCR investment in the Magellan cloud infrastructure

- Open Stack Cloud @ Argonne
- Open Stack Cloud @ Oak Ridge
- Cluster system @ Berkeley
- Cluster system @ Brookhaven

Current configuration of 700 nodes homed at ANL optimized for heterogeneous applications

KBASE
predictive biology
DOE Systems Biology Knowledgebase

The Kbase Cloud Architecture



KBASE
predictive biology
DOE Systems Biology Knowledgebase

The KBase Services

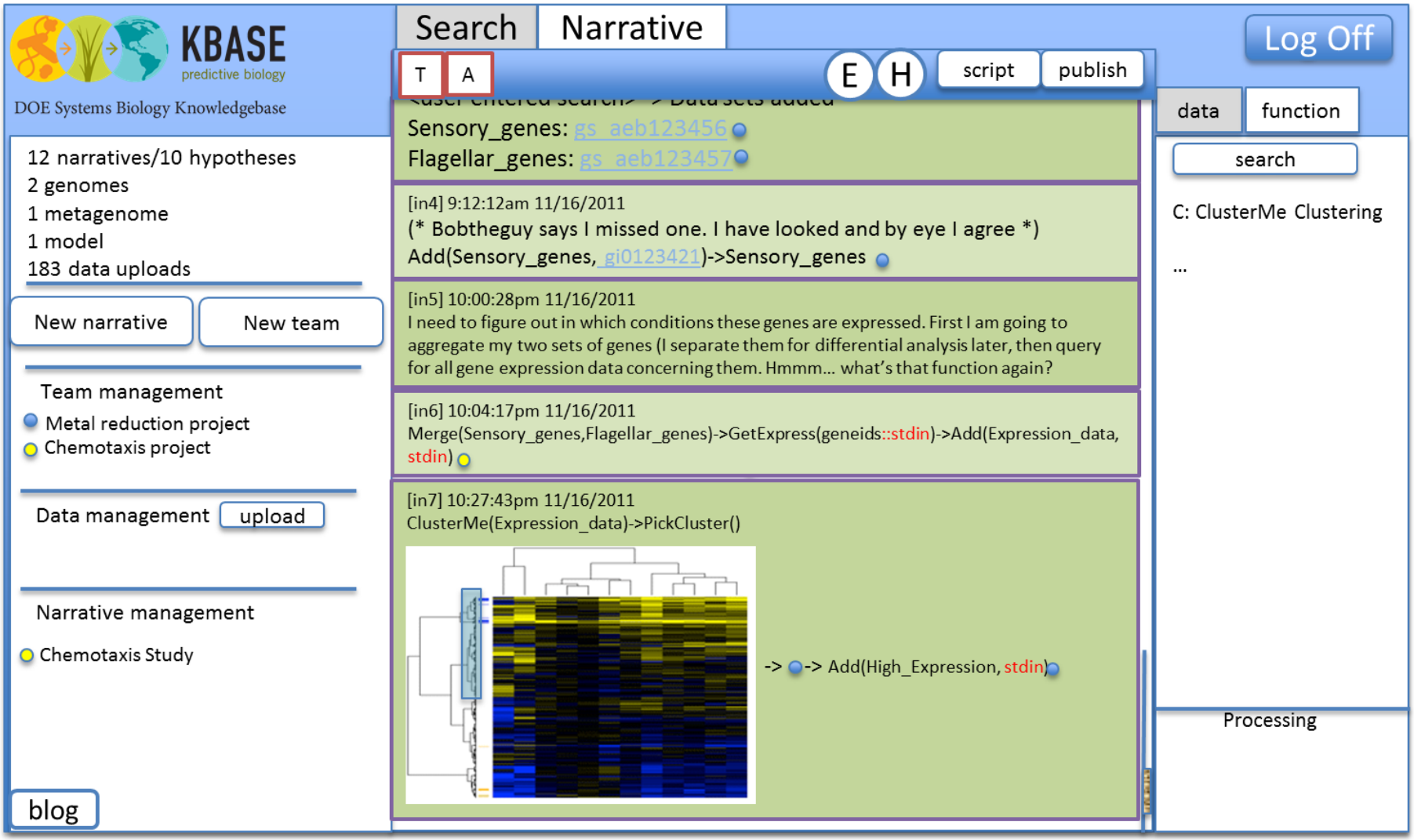
Services Oriented Architecture: The KBase Unified API access to a highly diverse set of services ranging from quick retrieval of simple data to massive computations on the KBase Cloud.

- In a SOA the system is functionally decomposed into many services each of which is implemented as one or more servers.
- Our long-term goal includes community developed and contributed services. Our initial set of services will be backed by the following example servers:

Genomic Servers	***Protein Family Servers***	***Phenotype Servers***	***Polymorphism Servers***
Compound and Reaction Data Servers	***Metabolic Modeling Servers***	***Expression Data Servers***	***Regulatory Models Servers***

KBASE predictive biology
DOE Systems Biology Knowledgebase

Concept: KBase User Experience



KBASE
predictive biology
DOE Systems Biology Knowledgebase

Development Schedule

A series of system builds occurring every quarter will enable a graded process. Successive builds will expand community involvement.

Feb 2012 – Development release (internal target)

- debug release engineering, prototype deployments, initial data models and data loads, non-unified API, performance testing, architecture refinement

May 2012 – Alpha release (internal target, invited testing)

- draft tutorials, v0.0 database loads, draft API (performance and ubiquity unified prototypes), draft UI library, domain workflow drafts, cloud and cluster services

You Are Here

Aug 2012 – Beta Release (early adopter beta testing)

- workflow function complete, API refinement, v1.0 database loads, prototype plug-in interfaces, prototype galaxy support, performance debugging

Nov 2012 – Production Release Candidate (public beta testing)

- draft website, draft documentation, full functionality API, draft UI v1.0, database loads v2.0, significant number of external beta test users

Feb 2013 – KBase Production Release

- public website, unified API, initial production UI, database loads v3.0 (microbes, community, plant databases), production demonstration workflows, and replication